FEBS Letters 584 (2010) 4129–4130

FEBS Letters

journal homepage: www.FEBSLetters.org

Editorial

The FEBS Letters SDA corpus: A collection of protein interaction articles with high quality annotations for the BioCreative II.5 online challenge and the text mining community

Establishing the interactome, a simplified but complete representation of the functional protein mesh, is a prerequisite for any attempt to model cell physiology [1,2]. Protein interaction databases such as MINT [3,4], IntAct [5], DIP [6], or BioGRID [7] strive to recapture the information published in a human readable format in scientific journals and to organize it in a structure that can be understood by a computer. However, this process is time-consuming and databases cannot keep up with the steadily growing amount of protein interaction information published in the scientific literature [8].

As an approach to relieve this problem, in 2008 FEBS Letters started to append so called Structured Digital Abstracts (SDA) to the "traditional" abstracts, summarizing the protein interaction information reported in the article in a structured format that makes use of a controlled vocabulary and of a tight syntactical convention [8]. SDAs are both human- and machine-readable and describe the interaction partners, via their UniProt identifiers ("accessions") [9]. In addition, they include information about the interaction type and the experimental method using the PSI MI controlled vocabulary [10]. To produce these SDAs, FEBS Letters asked authors to provide the necessary annotations on a voluntary basis during most of 2008. MINT curators refined these data in an interactive dialog with the authors to ensure a high quality standard [11] of these annotations.

We have recently evaluated the possibility of facilitating this author and curator task by introducing text-mining tools in the SDA generation process. Over the past decade, a broad range of text-mining and information extraction approaches have been developed to address the issues of annotating biologically relevant text and of extraction of information such as protein-protein interactions [12]. The BioCreative challenge (Critical Assessment of Information Extraction in Biology, see www.biocreative.org) consists of a collaborative initiative to provide a common evaluation framework for monitoring and assessing the state-of-the-art of text-mining systems applied to biologically relevant problems, similar to the CASP challenges for protein structure prediction [13]. So far, there have been three such community efforts: BioCreative I, in 2004 [14], BioCreative II in 2007 [15], and BioCreative II.5 in 2009 [16,17]. A fourth challenge, BioCreative III, has begun in 2010.

The last two BioCreative challenges (II and II.5) focused on protein-protein interaction extraction [18]. The two main tasks for the automated systems were (i) extraction of UniProt accessions (database identifiers) for proteins that have experimental interaction evidence descriptions in the body of the articles and (ii) identification of a list of interacting protein pairs. To this end, text-mining challenges require a collection of relevant annotated texts (called a "corpus") that can subsequently be used by automated machine learning systems to produce the corresponding annotations. Traditionally, the availability of a sufficiently sized corpus together with a set of relevant annotations on that corpus (called "ground truth", or "gold standard") has been one of the main hurdles for text-mining. Since BioCreative II, the MINT and IntAct [19] databases have contributed their curation facilities to provide these high-accuracy annotations. The corresponding annotated corpus represents a lasting asset for the text mining community since these freely available corpora allow researchers to further improve systems and methods long after the challenges.

With this background, for BioCreative II.5, a unique setting was created: In conjunction with FEBS Letters and MINT, the BioCreative II.5 organizers (Alfonso Valencia, Gianni Cesareni, Lynette Hirschman, Scott A. Mardis, Martin Krallinger, and Florian Leitner) announced a challenge to the biological text-mining community, asking them to reproduce these annotations with their automated systems in a realistic, online setting – i.e., by providing web-servers that could reproduce these annotations upon request within limited time constraints. However, to hold such a challenge, a reasonably sized corpus is required that can be distributed to the participants. FEBS provided the rights to distribute two years worth of FEBS Letters publications to the participants, in total 1190 articles, in machine-readable format (XML). In addition, MINT made an additional effort to provide high-quality annotations for the 122 protein interaction articles in this set including those articles authors annotated during the FEBS Letters experiment. The remaining 1068 articles that did not contain protein interaction information were used as negative examples for training machine-learning systems to discern relevant from irrelevant papers; a task database curators carry out manually, making a decision whether to annotate an article or not.

The BioCreative II.5 challenge was held in spring 2009, with a subsequent workshop in October, in Madrid, Spain. The three tasks were (1) article classification (annotation-relevant or not), (2) UniProt identifier assignment (identifying the relevant, interacting proteins), and (3) the extraction of the actual binary interaction pairs. As expected, the performance of automated systems did not match that of human experts (curators). However, the performance of systems was significantly higher than anticipated and was deemed sufficiently accurate to assist human annotators in tasks such as identifying the articles relevant for annotation, or reducing the time-consuming task of finding the correct database

doi:10.1016/j.febslet.2010.08.026

identifiers. The results of the best text mining systems were of sufficient quality to make it possible to consider integrating them in the SDA production process. Details of this collaborative effort between FEBS Letters, MINT, and BioCreative are published independently in Nature Biotechnology [16], while a special issue on the challenge itself, covering the technical aspects and participants' system descriptions, is being published in ACM/IEEE Transactions on Computational Biology and Bioinformatics [17].

The BioCreative II.5 machine-readable corpus of nearly 1200 FEBS Letters articles provided by FEBS and Elsevier is freely accessible to computational biologists, and together with the MINT and BioCreative annotations it provides a large resource of intrinsic scientific value. Both, the corpus and more information about the BioCreative resources, can be found on the BioCreative website at http://www.biocreative.org/; to download the corpus and other data, visit the Resources section. (Note that downloading resources requires users to create an account and accept the terms of using the data provided exclusively for scientific purposes.)

SDAs are now systematically added to all FEBS Letters manuscripts that contain protein-protein interactions. In addition, in 2009, FEBS Journal also started publishing manuscripts with SDAs bringing added value to their manuscripts as well. We now look forward to integrate the text-mining pipeline [20] in the SDA production pipeline in collaboration with text-mining groups, authors, publishing/editorial houses, journals and databases. The BioCreative III experiment will directly address this point by assessing the status and possibilities of the interaction between human and the text-mining systems incorporated in appropriate interfaces (see http://www.biocreative.org/news/chapter/biocreative-iii/).

References

[1] Sharan, R. and Ideker, T. (2006) Modeling cellular machinery through biological network comparison. Nat. Biotechnol. 24, 427–433.
[2] Sharan, R., Ulitsky, I. and Shamir, R. (2007) Network-based prediction of protein function. Mol. Syst. Biol. 3, 88.
[3] Ceol, A., Chatr-aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L. and Cesareni, G. (2010) MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. 38, D532–D539.
[4] Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. and Cesareni, G. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res. 35, D572–D574.
[5] Aranda, B. et al (2010) The IntAct molecular interaction database in 2010. Nucleic Acids Res. 38, D525–D531.
[6] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 32, D449–D451.
[7] Breitkreutz, B.-J. et al (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. 36, D637–D640.
[8] Ceol, A., Chatr-aryamontri, A., Licata, L. and Cesareni, G. (2008) Linking entries in protein interaction database to structured text: the FEBS Letters experiment. FEBS Lett. 582, 1171–1177.
[9] UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 38, D142–D148.
[10] Hermjakob, H. et al (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. Nat. Biotechnol. 22, 177–183.
[11] Orchard, S. et al (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat. Biotechnol. 25, 894–898.
[12] Krallinger, M., Valencia, A. and Hirschman, L. (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. 9 (Suppl. 2), S8.
[13] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. and Tramontano, A. (2009) Critical assessment of methods of protein structure prediction – Round VIII. Proteins 77 (Suppl. 9), 1–4.
[14] Blaschke, C., Hirschman, L., Yeh, A. and Valencia, A. (2003) Critical assessment of information extraction systems in biology. Comp. Funct. Genomics 4, 674–677.
[15] Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L. and Valencia, A. (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol. 9 (Suppl. 2), S1.
[16] Leitner, F. et al. (2010) The FEBS Letters/BioCreative II.5 Experiment: Making Biological Information Accessible. Nat. Biotechnol., in press.
[17] Leitner, F., Mardis, S.A., Krallinger, M., Cesareni, G., Hirschman, L. and Valencia, A. (2010) An overview of BioCreative II 5. IEEE-ACM T Comput. Biol. 7 (3), 385–399.
[18] Krallinger, M., Leitner, F., Rodriguez-Penagos, C. and Valencia, A. (2008) Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biol. 9 (Suppl. 2), S4.
[19] Chatr-aryamontri, A. et al (2008) MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. Genome Biol. 9 (Suppl. 2), S5.
[20] Leitner, F. et al (2008) Introducing meta-services for biomedical information extraction. Genome Biol. 9 (Suppl. 2), S6.

Florian Leitner
Martin Krallinger
Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Gianni Cesareni
Department of Biology, University of Rome Tor Vergata, Rome, Italy

Istituto Ricovero e Cura a Carattere Scientifico, Fondazione Santa Lucia, Rome, Italy

Alfonso Valencia
Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
E-mail address: valencia@cnio.es

Available online 20 August 2010